Projektausschreibung: Themenfeld "Nutzung von Large Language Models (LLMs) in Unternehmen, Architektur und Engineering für Retrieval Augmented Generation (RAG), Model Context Protocol (MCP) und agentische Systeme"

1. Projekt

1.1 Titel

Enterprise-Integration von LLMs: Retrieval-Augmented Generation, MCP-basierte Werkzeugnutzung und agentische Systeme im Unternehmenskontext

1.2 Laufzeit, Mittel (Höhe, Mittelgeber), Einbindung in größeres Projekt

3 Semester, Eigenfinanzierung

1.3 Kurzbeschreibung der Ziele / Aufgaben

Unternehmen wollen Large Language Models (LLMs) zunehmend auf ihre eigenen Daten, Prozesse und Systeme anwenden. Der Schritt vom generischen Chat-Modell zur verlässlichen, unternehmensweiten Anwendung erfordert allerdings mehr als reine Modellqualität: Er erfordert eine belastbare Architektur, die Datenzugriff, Werkzeugnutzung und Handlungsfähigkeit kontrolliert mit dem Modell verbindet.

Drei Bausteine stehen hierbei im Zentrum:

- **Retrieval-Augmented Generation (RAG):** Verknüpfung des Modells mit unternehmenseigenen Datenquellen (z. B. Dokumente, Wikis, Datenbanken) unter Berücksichtigung von Datenschutz und Zugriffsrechten.
- **Model Context Protocol (MCP):** Standardisierte Anbindung von Werkzeugen, Daten und Fachsystemen an LLMs, um diese kontrolliert und wiederverwendbar in Unternehmensprozesse einzubinden.
- **Agentische Systeme:** Orchestrierung mehrerer Werkzeug- und Entscheidungsschritte zu Agenten, die komplexere Aufgaben im Unternehmen eigenständig bearbeiten.

Ziel des Forschungsprojekts ist es, eine Referenzarchitektur für die Enterprise-Integration von LLMs entlang dieser drei Bausteine aufzubauen, in repräsentativen Szenarien zu evaluieren und daraus wissenschaftlich fundierte Architektur- und Engineering-Muster abzuleiten.

1.4 Wissenschaftlicher Anteil für Forschungsmaster

Das Forschungsprojekt kann je nach Interessenslage unterschiedliche Schwerpunkte setzen. Aufbauend auf bestehenden Vorarbeiten zu RAG- und LLM-Deployment im Forschungscluster sind folgende Forschungsthemen möglich:

1. **Enterprise-RAG-Architekturen:** Entwurf und Evaluation von RAG-Pipelines (Retrieval, Ranking, Generation) unter realistischen Unternehmensbedingungen, inklusive rollenbasierter Zugriffskontrolle (RBAC) und On-Premises-Anforderungen.
2. **Hybride Retrieval-Strategien:** Kombination aus lexikalischem (BM25) und semantischem Retrieval mit Re-Ranking, evaluiert auf unternehmenstypischen Dokumentkorpora.
3. **MCP-basierte Werkzeug- und Datenanbindung:** Entwurf und prototypische Umsetzung von MCP-Servern zur kontrollierten Anbindung von Fachsystemen, Datenquellen und Werkzeugen an LLMs; Bewertung von Sicherheits-, Governance- und Integrationsaspekten.
4. **Agentische Systeme im Unternehmenskontext:** Vergleich von Planungs-, Reflection- und Tool-Use-Mustern für LLM-basierte Agenten; Evaluation auf typischen Aufgaben (z. B. Informationsbeschaffung, Prozessautomatisierung, Assistenz).
5. **Betriebs- und Qualitätsaspekte:** Messung und Steuerung von Qualität, Latenz, Kosten und Risiken (z. B. Halluzinationen, Datenabfluss) im produktiven Einsatz.
6. **Architekturforschung für LLM-basierte Enterprise-Systeme:** Ableitung wiederverwendbarer Architektur- und Integrations-Muster.

---

2. Durchführende Stelle

2.1 Institut / Labor / Arbeitsplatz

Hochschule Ansbach, Hochschule Ansbach, CCS – Center for Signal Analysis of Complex Systems

2.2 Betreuerin / Co-Betreuer / Betreuende Laboringenieurinnen

Prof. Nicolas Weeger

2.3 Notwendige Ausstattung vorhanden / wird in Projektlaufzeit beschafft

Die notwendige Ausstattung ist überwiegend vorhanden: GPU-Workstation mit NVIDIA RTX 5090 sowie Zugang zu einer DGX-Spark-Plattform (GB10 Blackwell, Unified Memory) für lokale LLM-Deployments (z. B. über vLLM). Container- und Integrationsumgebung ist etabliert. Für Vergleichsmessungen kann zusätzlich auf eine performantere Hardware zurückgegriffen werden.

---

3. Reporting

3.1 Rahmen für Projekt- / Masterseminar vorhanden

Projektfortschritt und Ergebnisse sind im Rahmen der APR-Seminare und Tagungen darzustellen.

3.2 Veröffentlichung geplant auf Konferenz / in Zeitschrift / als Patentanmeldung

Veröffentlichungen sind entsprechend dem Projektfortschritt vorgesehen, u. a. auf Konferenzen im Bereich Software-Architektur und Software Engineering (ICSA, ECSA, ICSE) sowie in peer-reviewed

Zeitschriften (z. B. TSE, TOSEM, IEEE Software, Software: Practice and Experience). Zusätzlich sind Beiträge auf anwendungsorientierten Formaten wie möglich.

---

4. Anforderungen an Bewerber

4.1 Gewünschte/vorausgesetzte Fachrichtung eines Hochschulabschlusses

Bachelor of Science oder Bachelor of Engineering in Informatik, Wirtschaftsinformatik, KI, Data Science, Software Engineering oder verwandten Disziplinen. Interesse an der Integration von KI-Systemen in Unternehmensarchitekturen sowie die Bereitschaft, sich in neue Themenfelder einzuarbeiten.

4.2 Vorteilhaft sind folgende Vertiefungen / praktische Erfahrungen

- Solide Python-Kenntnisse
- Grundkenntnisse im Umgang mit LLMs / Transformern und typischen Frameworks
- Erfahrung oder starkes Interesse an Software-Architektur und Systemintegration (REST, Container, Authentifizierung / Autorisierung)
- Sicheres Englisch in Wort und Schrift (wissenschaftliche Publikationen)

Bei Interesse oder Fragen wenden Sie sich bitte an Prof. Nicolas Weeger (nicolas.weeger@hs-ansbach.de)